Las aplicaciones de la Lexicografía Histórica a la recuperación documental a través del léxico notarial aragonés del siglo XV

Mª Mercedes Muñoz Escolá
Universidad de Zaragoza

0.1 Resumen

Propuesta de desarrollo de un sistema de información histórica para el tratamiento y recuperación de documentos notariales aragoneses del siglo XV, que constaría de una base de datos textual y una léxica. La base de datos léxica abordaría el análisis de las formas léxicas aragonesas recogidas en las fuentes. Finalmente, se estudian los potenciales usuarios de dicho sistema de información.

Palabras clave: Sistemas de información histórica. Documentación notarial aragonesa. Siglo XV.

0.2. Abstract

Proposal of a historical information system for treating and retrieving 15th century Aragonian notarial documents. The system would consist of two databases: a textual one and a lexical one. The lexical database would store the lexical analysis and parsing of the Aragonian-language forms documented in the historical sources. Finally, the possible users of the system are considered.

Keywords: Historical information systems. Aragonian notarial documents. 15th century.

1. Introducción

"Un sistema de información histórica debe ser definido como un sistema automatizado integrado por un conjunto de bases de datos y procedimientos formales cuyo diseño y mantenimiento sirva para almacenar, tratar y recuperar información histórica". Esta definición que ofrece García Marco (1993) es sus-

ceptible de ser aplicada a cualquier sistema de información añadiendo que debe servir al mayor número de usuarios posible.

La propuesta de Sistema de información que se ofrece a continuación responde al objetivo antes expuesto y pretende llegar a un colectivo múltiple integrado principalmente por lingüistas, historiadores, profesionales del derecho en su vertiente histórica y/o de Derecho Foral aragonés y sociólogos de la Historia.

El proyecto supone la creación de una base de datos textual que contenga documentos notariales aragoneses del siglo XV y que a su vez esté relacionada con otra base de datos de tipo léxico que aborde el estudio y análisis de las formas léxicas aragonesas recogidas en los documentos primarios o fuentes.

2. Diseño de la base de datos textual

Para el estudioso de la lengua o del hecho histórico el acceso al documento primario es siempre importante y en algunos casos se hace imprescindible. El problema con el que se encuentran ambos es la dificultad de acceso a los mismos debido a su dispersión y en algunos casos a su mala conservación. Sabido es la cantidad de documentación de notable interés perdida a lo largo del tiempo unas veces por causas no controlables y las más por desidia y abandono.

Actualmente esta dificultad ha quedado solucionada en parte por la tecnología actual que posibilita la digitalización de los documentos y por consiguiente el acceso a ellos con una mejora sustancial en su lectura. De hecho la técnica de tratamiento de imágenes permite de modo electrónico la "restauración" del documento haciendo legibles textos que parecían irrecuperables hasta ahora. Tal situación tecnológica nos lleva a plantear la creación de una base de datos textual en la que se recogerían los documentos originales digitalizados.

Para historiadores y lingüistas la documentación de los siglos bajo-medievales es rica y sugerente, pero buena parte de ella está inédita en archivos locales. Muchos estudiosos han subrayado lo positivo del trabajo lingüístico aplicado a documentación notarial, en especial a los *inventarios de bienes* que se realizaban a propósito de donaciones, herencias, embargos o matrimonio.

Son estas las razones para la elección de este tipo de documentación como base de datos textual.

A continuación pasamos a plantear el diseño de la base textual.

Al estar compuesta por un número de documentos que si bien pertenecen a la misma tipología documental *inventarios de bienes* son distintos en su localización y datación, se impone la creación de una serie de campos que sirvan de identificación del documento y lo singularicen dentro de la pluralidad. De tal manera cada registro se compondrá de:

- *Nº de identificación del documento:* Este tipo de numeración es la que proporciona el propio sistema cuando introducimos un nuevo registro. Para el usuario carece de aplicación práctica.
- *Código del documento:* Este código puede ser identificativo del documento para el usuario. En los textos literarios medievales lo constituye el título pero en un documento notarial puede ser un conjunto de letras y números significativo para el usuario como por ejemplo: la abreviatura del lugar en el que se otorga el documento más la del notario que lo firma seguido de un número correlativo que se asigna a los documentos de un mismo notario.
- *Tipología:* Como se ha visto la tipología documental es la misma para todos los documentos que integran la base pero dentro de esta tipología general que los aúna podemos crear una tipología específica que atienda al objeto o fin para el que ha sido realizado el inventario y así por ejemplo para las capitulaciones matrimoniales; herencias, etc. incluiríamos también lugar., fecha, notario, peticionario, número de folios, localización física, y el transcriptor, en cuyo campo se indica el nombre de la persona o personas que han realizado la transcripción paleográfica.

El conjunto de todos estos campos tratan de describir el documento e individualizarlo frente a los otros así como crear una serie de referencias de búsqueda.

A continuación de esta información identificativa aparecería el texto original digitalizado. La importancia que tiene el manejo del documento original ya se ha apuntado más arriba pero para poder aplicar tratamientos electrónicos a la información es necesaria la transcripción de los mismos.

La transcripción de los textos plantea dos problemas: el satisfacer las diferentes necesidades de los usuarios del sistema y la aplicación de una normalización.

En cuanto al primer problema, el estudio del tipo de usuarios que van a acceder y manejar la documentación aconseja un modelo de transcripción distinta para cada caso. La edición más cercana al texto físico es la llamada *paleográfica* que consiste en la transcripción del texto de la manera más fiel posible, manteniendo las grafías según aparecen en él aunque se sepa que son variantes gráficas y corresponden a un mismo fonema. Incluso se conserva la disposición del texto y los signos diacríticos. Este tipo de transcripción es la que más interesa a los lingüistas por no decir la única. Por el contrario para el historiador el mantenimiento de grafías no es vital y prefiere un tipo de transcripción mas cercana al texto moderno que facilite su lectura. En este tipo de transcripción hay una unificación de grafías que representan un mismo sonido. Es la que Marcos-Marín (1994) llama *fonemática.* Una edición más avanzada y ya al alcance de cualquier usuario es la *modernizada*, en la que las grafías antiguas se sustituyen por las modernas.

Los diferentes tipos de transcripción plantean problemas de normalización. La paleográfica es la que presenta mayores dificultades pero existe una serie de normas establecidas por el Hispanic Seminary of Medieval Studies (HSMS) de la Universidad de Wisconsin en Madison que pueden servir de referente pues ya han sido utilizadas en España por el programa ADMYTE (Archivo Digital de Manuscritos y Textos Españoles). El HSMS propone una serie de estándares de normalización y codificación utilizados por su grupo investigador en el proyecto del *Dictionary of the Old Spanich Language*.

La llamada transcripción fonemática se crea a partir de la anterior mediante la ejecución de un programa sencillo que normalice el texto eliminando las discrepancias gráficas.

En ambos casos el usuario debe saber qué normalización se ha seguido y desde la pantalla de transcripción se le debe proporcionar el acceso a los estándares utilizados.

Así mismo es preciso al crear la base de datos correlacionar en todo momento el o los textos transcritos con el texto digital ya que el paso de uno a otro debe hacerse de forma inmediata, incluso debe existir la posibilidad de traer a pantalla el documento digital y el transcrito al mismo tiempo para poder confrontarlos. Naturalmente todo ello tiene que ser transparente para el usuario a través de un interfaz amigable.

3. El diseño de la base léxica

El objetivo que se persigue con la creación de una base de datos léxica es doble. Por un lado, crear dos listas de palabras que posteriormente formarán parte del Glosario. Por otro, el análisis léxico de los términos aragoneses que aparecen en los textos.

En cuanto a las listas que forman el Glosario, la primera de ellas—*Lista de Formas*—contendrá todas las palabras de la base incluyendo las variantes gráficas de una misma palabra. La segunda—*Lista de Lemas*—estará compuesta por las palabras que, después de haber pasado por un proceso de lematización, consideremos que deberán ser la entrada léxica propiamente dicha. El proceso de lematización constituye la agrupación de las variantes gráficas, fonéticas y morfológicas medievales que comparten una misma base léxica. Por lo tanto a través de la palabra que se erija como lema accederemos a todas las variantes que existen del término.

Ambas listas son decisivas en el momento de la búsqueda y recuperación de los documentos. Cada una de ellas está orientada a un tipo de usuarios. Mientras que la primera tendrá mayor importancia para lingüistas y filólogos, la segunda lo será para el resto de investigadores ya que el usuario no tendrá que ir buscan-

do en la lista de formas palabras que ocupen distintas posiciones en la misma sino que accederá a todas las variantes desde una sola palabra o lema.

El Glosario nos ofrecerá varias opciones de selección:

- Acceso a los documentos que contengan la forma o lema marcados. La presentación del texto será bien facsímil, bien transcrito en cualquiera de sus variantes o bien facsímil y transcrito al mismo tiempo a fin de confrontar ambos.
- Ver la palabra clave dentro de un contexto. Es decir, ver la parte del texto del o de los documentos en la que aparece la palabra. El número de líneas del contexto que deseamos que aparezca puede ser modificable por el usuario desde una sola línea a varias anteriores y posteriores a la palabra marcada.
- Ver la distribución de frecuencia de la palabra marcada. Este punto es importante para su aplicación en estadísticas léxicas.

El análisis de las voces aragonesas nos permitirá crear un fondo susceptible de ser utilizado en un futuro *Diccionario Histórico Aragonés*, proyecto largamente perseguido por el Departamento de Lingüística Española de la Facultad de Filosofía y Letras de Zaragoza.

Naturalmente el estudio y análisis de este léxico supone la aprobación de unas normas de codificación y presentación de cada entrada léxica. Las normas de codificación vendrán impuestas por el programa que se utilice mientras que la presentación de la entrada tendrá que ser discutida y aprobada por el equipo de lingüistas que se dediquen al estudio del léxico. En todo caso las entradas léxicas coincidirán con los lemas de la Lista de Lemas ya mencionada. De esta forma todas las variantes de una palabra quedarán agrupadas bajo un mismo término.

4. La búsqueda y recuperación

El Sistema deberá permitir la búsqueda desde el Glosario de palabras, lemas o formas, y se efectuará sobre toda la base de datos textual o sólo sobre los documentos señalados en una selección anterior. Podremos seleccionar el tipo de inventario de bienes sobre el que deseamos trabajar o bien por medio del lenguaje de interrogación del programa seleccionar los documentos por cualquiera de los campos que contiene cada registro.

Como ya se ha mencionado más arriba el Glosario nos proporciona otras opciones: ver la palabra en contexto y ver su distribución.

Desde el documento recuperado, marcando una determinada palabra, el Sistema deberá permitirnos el acceso a la base de datos léxica: al Diccionario si

lo que queremos es obtener información sobre el término y su significado o a cualquiera de las listas del Glosario.

Así mismo desde cualquier texto recuperado se deberá permitir el acceso al mismo texto en cualquiera de sus variantes de transcripción o facsímil.

5.Conclusiones

El proyecto aquí presentado, aunque muy específico y limitado en espacio, tiempo y temática, pretende animar al trabajo multidisciplinar y aunar esfuerzos en la creación de Sistemas de Información que sirvan al mayor número de usuarios posible. Cualquier sistema de información por pequeño que éste sea conlleva un ahorro de energía y esfuerzos en futuras investigaciones y puede servir de base a la creación de otros de mayor envergadura.

Es a través de estos pequeños y localizados proyectos de sistemas de información como podremos llegar a realizar grandes obras para las que no hay presupuesto y están paralizadas desde hace años.

6.Referencias

Buesa, T. (1989). Estudios filológicos aragoneses. Zaragoza : Universidad, 1989.

Fort Cañellas, M.R. (1994). Léxico romance en los documentos medievales aragoneses (s. XI y XII). Zaragoza : Diputación General de Aragón, 1994.

García Marco, F.J. (1994). Knowledge organisation in historical information sistems. // Advances in Knowledge Organisation. 4 (1994). 81-90.

Lagüéns Gracia, V. (1992). Léxico jurídico en documentos aragoneses de la Edad Media (s. XIV y XV). Zaragoza : Diputación General de Aragón, 1992.

Mackenzie, D. (1984). A Manual of Manuscript Transcription for the Dictionary of the old Spanich Language (With Spanich traslation by José Luis Moure). Madison : Hispanic Seminary of Medieval Studies, 1984.

Marcos Marín, F.A. (1994). Informática y humanidades. Madrid : Gredos, 1994.

Sesma Muñoz, A.; Líbano Zumalacárregui, A (1982). Léxico del comercio medieval en Aragón (s. XV). Zaragoza : Institución Fernando el Católico, 1982.